Pstat 131 Homework 1

Yu Tian

Spring 2022-04-10

Machine Learning Main Ideas

Question 1:

Define supervised and unsupervised learning. What are the difference(s) between them?

Answer (from the lecture/page #26 of book)

Supervised learning:

For each observation of the predictor measurement(s) x_i, $i = 1, ..., n$ there is an associated response measurement y_i. Then fit a model that relates the response to the predictors to accurately predict the response for future observations (prediction) or better understanding the relationship between the response and the predictors (inference). Thus, supervised learning has a supervisor, and it need to give the model observed output and input. The specific learning process is prediction (accurately predict future response given predictors), estimation (understand how predictors affect response), model selection (find the "best" model for response given predictors), inference (assess the quality of our predictions and (or) estimation).

Many statistical learning method operate in the supervised learning, like linear regression, logistic regression, k-nearest neighbors, decision trees, random forests, support vector machine(s), neural network.

Unsupervised learning:

For every observation $i = 1, ..., n$, we observe a vector of measurements x_i but no associated response y_i. In this setting, we are in some sense working blind since we lack a response variable that can supervise our analysis.We can use it to understand the relationships between the variables or between the observations. Thus, unsupervised learning learn without a supervisor.

Some statistical learning method operate in the unsupervised learning, like Principal Component Analysis(PCA), k-means clustering, Hierarchical clustering, neural networks.

The difference is that supervised learning with known response, but unsupervised learning with unknown response.

Question 2:

Explain the difference between a regression model and a classification model, specifically in the context of machine learning.

Answer (from the lecture/page #28 of book)

Regression model predicts a continuous outcome with a quantitative response Y, and quantitative variables take on numerical values.

Classification model predicts discrete class labels with a qualitative response Y, and qualitative variables take on values in one of K different classes, or categories.

Question 3:

Name two commonly used metrics for regression ML problems. Name two commonly used metrics for classification ML problems.

Answer (from the lecture)

Regression ML problems: Mean Squared Error (MSE) and Mean Absolute Error(MAE)

Classification ML problems: Accuracy and log-loss

Question 4:

As discussed, statistical models can be used for different purposes. These purposes can generally be classified into the following three categories. Provide a brief description of each.

Answer (from the lecture)

Descriptive models: Descriptive models best visually emphasize a trend in data, like using a line on a scatterplot.

Predictive models: Predictive models predict response (Y) with minimum reducible error without focusing on hypothesis tests.

Inferential models: Inferential models is significant to test theories (possibly) causal claims. They state relationship between outcome & predictors.

Question 5:

Predictive models are frequently used in machine learning, and they can usually be described as either mechanistic or empirically-driven. Answer the following questions.

a) Define mechanistic. Define empirically-driven. How do these model types differ? How are they similar?

b) In general, is a mechanistic or empirically-driven model easier to understand? Explain your choice.

c) Describe how the bias-variance tradeoff is related to the use of mechanistic or empirically-driven models.

Answer (from the lecture)

a) Mechanistic is parametric. For mechanistic predictive models, we assume a parametric form for $f(i.e., \beta_0 + \beta_1 + \ldots$. It won't match true unknown f. Also, It can be added parameter to be more flexibility. However, with too many parameter, it will be overfitting.

Empirically-driven is non-parametric. There is no assumptions about f. Empirically-driven predictive models requires a larger number of observations. It can be much more flexible by default. It also will have the situation of overfitting.

The difference is that mechanistic models are parametric, but the empirically-driven models are non-parametric. The simility is that they will have the possibility of being overfitting. They all can be more flexibility, but mechanistic models change it with the number of parameters, and empirically-driven models change it by default.

b) Mechanistic models will be easier to understand.

Explanation: We can predict results about one event in the society through collecting information and applying them to a theory. Only using the parametric form and adding parameters will be easier to understand.

c) The mechanistic models usually are more flexible, which means smaller bias and larger variance. The empirically-driven models relatively will have larger bias and smaller variance. Thus, the bias-variance tradeoff is related to the choice of mechanistic or empirically-driven models.

Question 6:

A political candidate's campaign has collected some detailed voter history data from their constituents. The campaign is interested in two questions:

Given a voter's profile/data, how likely is it that they will vote in favor of the candidate?

How would a voter's likelihood of support for the candidate change if they had personal contact with the candidate?

Classify each question as either predictive or inferential. Explain your reasoning for each.

Answer (according to the lecture)

"Given a voter's profile/data, how likely is it that they will vote in favor of the candidate?" is predictive. Since this question has given the actual data of observations to predict the response (the likelihood of voting in favor of the candidate), and it needs the accurate results with minimum reducible error, so this question is a predictive question.

"How would a voter's likelihood of support for the candidate change if they had personal contact with the candidate?" is inferential. This question is a causal hypothesis test,and it need to state a relationship inference according to the "voter's likelihood of support for the candidate" and "personal contact with the candidate".

Exploratory Data Analysis

```
mpg %>%
  head()
```

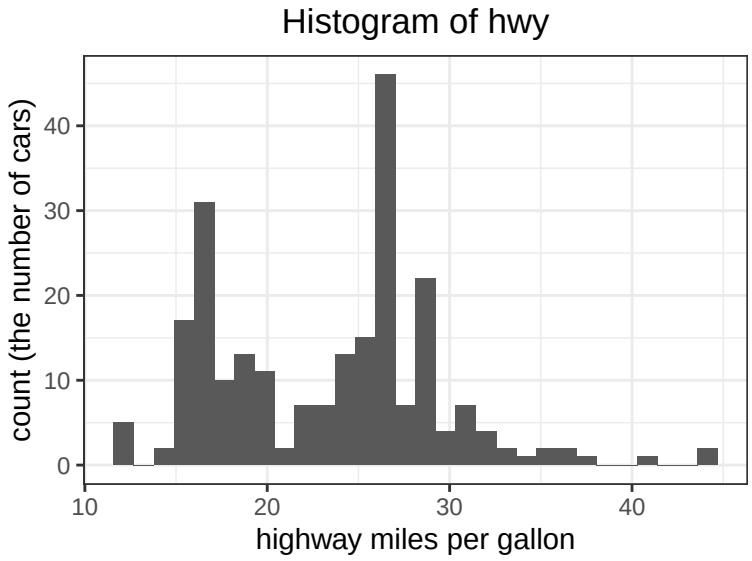
```
## # A tibble: 6 x 11
##   manufacturer model displ  year   cyl trans      drv     cty   hwy fl    class
##   <chr>        <chr> <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
## 1 audi         a4      1.8  1999     4 auto(l5)   f        18    29 p     compa~
## 2 audi         a4      1.8  1999     4 manual(m5) f        21    29 p     compa~
## 3 audi         a4      2    2008     4 manual(m6) f        20    31 p     compa~
## 4 audi         a4      2    2008     4 auto(av)   f        21    30 p     compa~
## 5 audi         a4      2.8  1999     6 auto(l5)   f        16    26 p     compa~
## 6 audi         a4      2.8  1999     6 manual(m5) f        18    26 p     compa~
```

Exercise 1:

We are interested in highway miles per gallon, or the hwy variable. Create a histogram of this variable. Describe what you see/learn.

```
# plot a histogram of highway miles per gallan
mpg %>%
  ggplot(aes(x=hwy)) +
  geom_histogram(bins=30)+
  labs(title="Histogram of hwy",
       x ="highway miles per gallon", y = "count (the number of cars)") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```

Answer



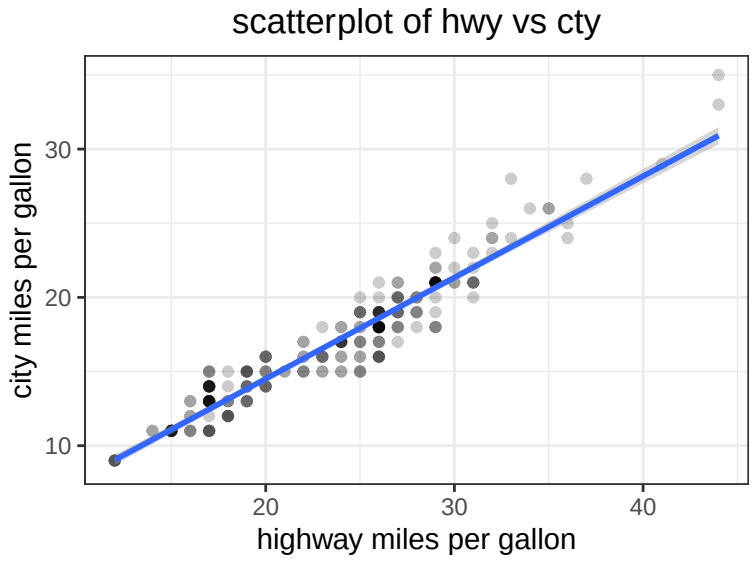
From the histogram, we can find that highway miles per gallon for all cars are between the range of 10 to 45. Most cars' hwy between 15 and 30. There are two peaks around 16 hwy and 26 hwy. There is a few cars has highway miles per gallon which more than 40. Also, the distribution is slight skewed.

Exercise 2:

Create a scatterplot. Put hwy on the x-axis and cty on the y-axis. Describe what you notice. Is there a relationship between hwy and cty? What does this mean?

```
# create a scatterplot with hwy on the x-axis and cty on the y-axis
mpg %>%
  ggplot(aes(x=hwy, y=cty)) +
  geom_point(alpha = 0.2) +
  labs(title = 'scatterplot of hwy vs cty',
       x='highway miles per gallon', y='city miles per gallon') +
  geom_smooth(method = 'lm', formula = 'y ~ x') +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```

Answer



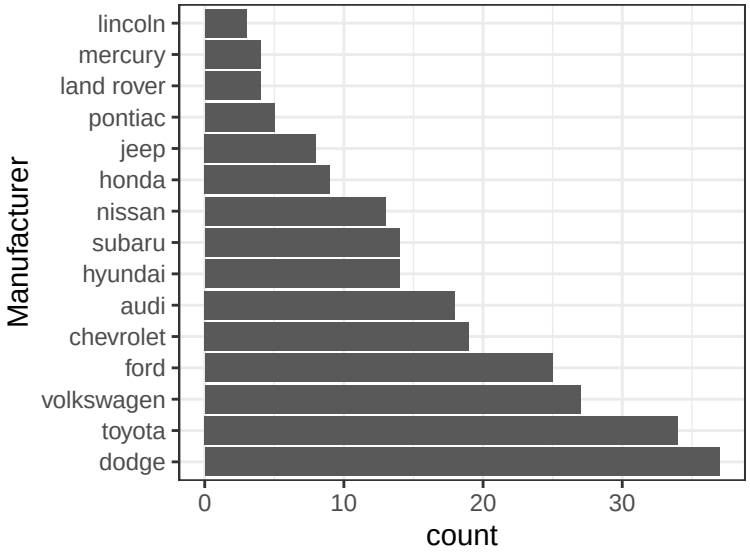
hwy and cty has a positive relationship and the spread of scatter around the blue line is almost even. As highway miles per gallon increasing, city miles per gallon will also increase.

Exercise 3:

Make a bar plot of manufacturer. Flip it so that the manufacturers are on the y-axis. Order the bars by height. Which manufacturer produced the most cars? Which produced the least?

```
# make a bar plot of manufacturers
mpg %>%
  ggplot(aes(x=reorder(manufacturer,rep(-1,length(manufacturer)),sum))) +
  geom_bar() +
  labs(x="Manufacturer") +
  coord_flip() +
  theme_bw()
```

Answer



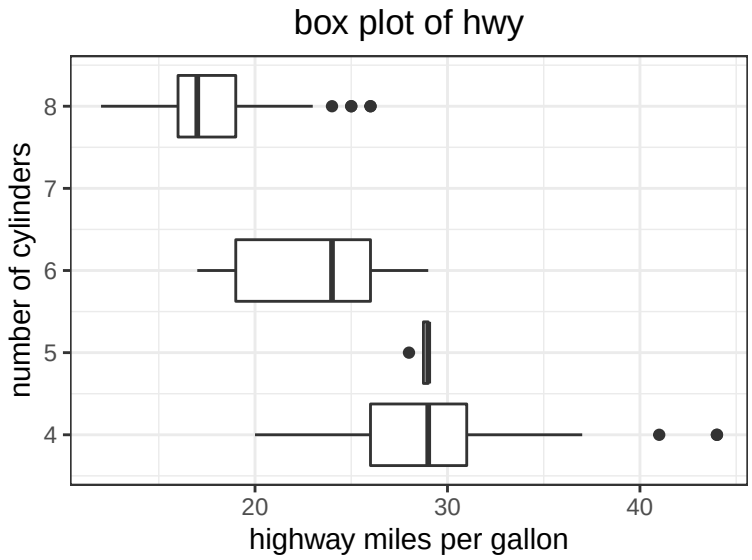
Dodge produced the most cars. Lincoln produced the least cars.

Exercise 4:

Make a box plot of hwy, grouped by cyl. Do you see a pattern? If so, what?

```
# make a box plot of hwy
mpg %>%
  ggplot(aes(x = hwy, y = cyl, group = cyl)) +
  geom_boxplot() +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))+
  labs(title = "box plot of hwy", x = "highway miles per gallon", y = "number of cylinders")
```

Answer



Yes, it has a patter. There is a negative relationship between hwy and cyl. With the higher number of cylinders, the values of highway miles per gallon is lower. They have a negative relationship.

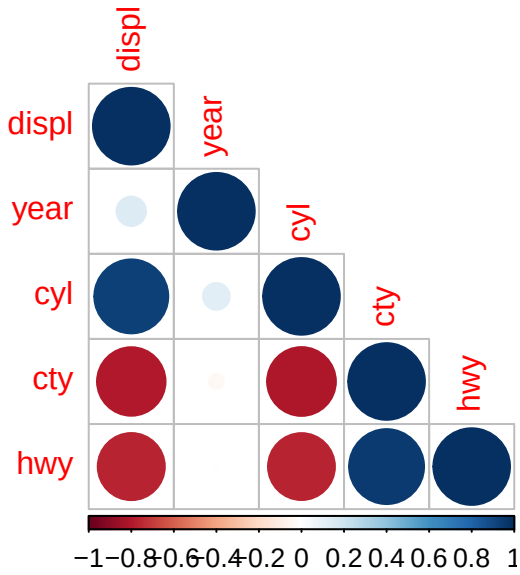
Exercise 5:

Use the corrplot package to make a lower triangle correlation matrix of the mpg dataset.

Which variables are positively or negatively correlated with which others? Do these relationships make sense to you? Are there any that surprise you?

```
# make a lower triangle correlation matrix of the mpg dataset
mpg %>%
  select(where(is.numeric)) %>%
  cor() %>%
  corrplot(type = 'lower', method = 'circle')
```

Answer



year is positively correlated with displ

cyl is positively correlated with displ and year

cty is negatively correlated with displ, year, and cyl

hwy is negatively correlated with displ, and cyl

hwy is positively correlated with cty

Yes, these relationships make sense to me, so there is no relationship that surprise me.